SPAM DETECTION IN COLLABORATIVE TAGGING

Bachelor of Technology

In

Computer Science and Engineering

By-

Dayadru Nayak
110CS0138

Manas Ranjan Sahoo
110CS0146

Department of Computer Science and Engineering
National Institute of Technology, Rourkela,
Sundargarh, Odisha, 769008, India
May, 2014

SPAM DETECTION IN COLLABORATIVE TAGGING

A thesis is submitted in partial fulfillment of the requirements for the degree of Bachelor of technology

In

Computer Science and Engineering

By-

Dayadru Nayak
110CS0138

Manas Ranjan Sahoo
110CS0146

Under the guidance of:
Prof. R. K. Mohapatra

Department of Computer Science and Engineering
National Institute of Technology, Rourkela,
Sundargarh, Odisha, 769008, India

Department of Computer Science and Engineering
National Institute of Technology, Rourkela
Rourkela – 769008, Odisha, India

Certificate

This is to certify that the work in the thesis entitled "SPAM DETECTION IN COLLABORATIVE TAGGING" submitted by **Dayadru Nayak and Manas Ranjan Sahoo**, in partial fulfillments of the requirements for the award of Bachelor of Technology Degree in the department of Computer Science and Engineering at National Institute of Technology, Rourkela is a record of original and authentic research carried out by him under my supervision and guidance.

Place: NIT Rourkela
Date: May, 2014

Prof. R. K. Mohapatra
Dept. of Computer Science and Engineering
National Institute of Technology, Rourkela
Odisha-769008

ACKNOWLEDGEMENT

I would like to express my earnest gratitude to my guide **Prof. R. K. Mohapatra** for his exemplary guidance and professional help. His profound insights helped me to reach my goal in this project. His support led to the encouragement to pursue this research.

I am highly indebted to **Prof. Santanu Kumar Rath**, Head of Department of Computer Science Department for giving the opportunity and help required every time. I would also like to thank all the faculty staff for their support.

I am thankful to all my friends whose constant motivation compelled me to finish the work.

Dayadru nayak
110cs0138

Manas Ranjan Sahoo
110cs0146

ABSTRACT

The algorithm which we will be proposing here in our project will be able to identify the spammers and demote their ranks cocooning the users from their malicious intents and gives popular and relevant resources in a collaborative tagging system or in online dating sites, or any other online forum where there are discussions like quora, amazon feedbacks etc. by a suitable algorithm on lines of an existing one but with multifaceating dimensions as against them.

We have taken the assumption that there are two factors on which the virtuosity of a user with reference to a resource or a document depends on. First and foremost an expert should have a rich content resource in his repertoire and his dexterity to find good resources, however the paraphernalia for rich resource is virtuosity of users who tagged it. Secondly, an expert should be first to identify intriguing or riveting documents. We propose an algorithm is designed based on the above ideas.

CONTENTS

CHAPTER 1

INTRODUCTION

1.1 FOLKSONOMY

A folksonomy is a system of classification emaciated from the practice of collaboratively creating and translating tags to footnote and classifying content; this is known as collaborative tagging[1].

In general social discussion forums like quora or social networks erratically have a list of documents or taggers taking into account of their frequency and time of appearance in the forum.

That the virtuosity of a tagger with reference to a specific discussion or document depends on the following rationality[1]:

1. Mutual fortification between virtuosity of user and the richness of a resource
2. The expert is capacitate and qualified enough to discover and detect rich and riveting resources before others

1.1.1 DEFINITION

In folksonomy, F is a tuple F = (U, T, D, R), where U is a set of users, T a set of tags, D a set of documents, and (R, U, T, D) a set of bookmarks[1].

It represents a user u is giving a document D a tag T .R is the set of tags by user c and $t \in R$

1.2 WEB CRAWLER

A Web crawler is an internet bot that systematically browses the World Wide Web typically for the purpose of Web Indexing.

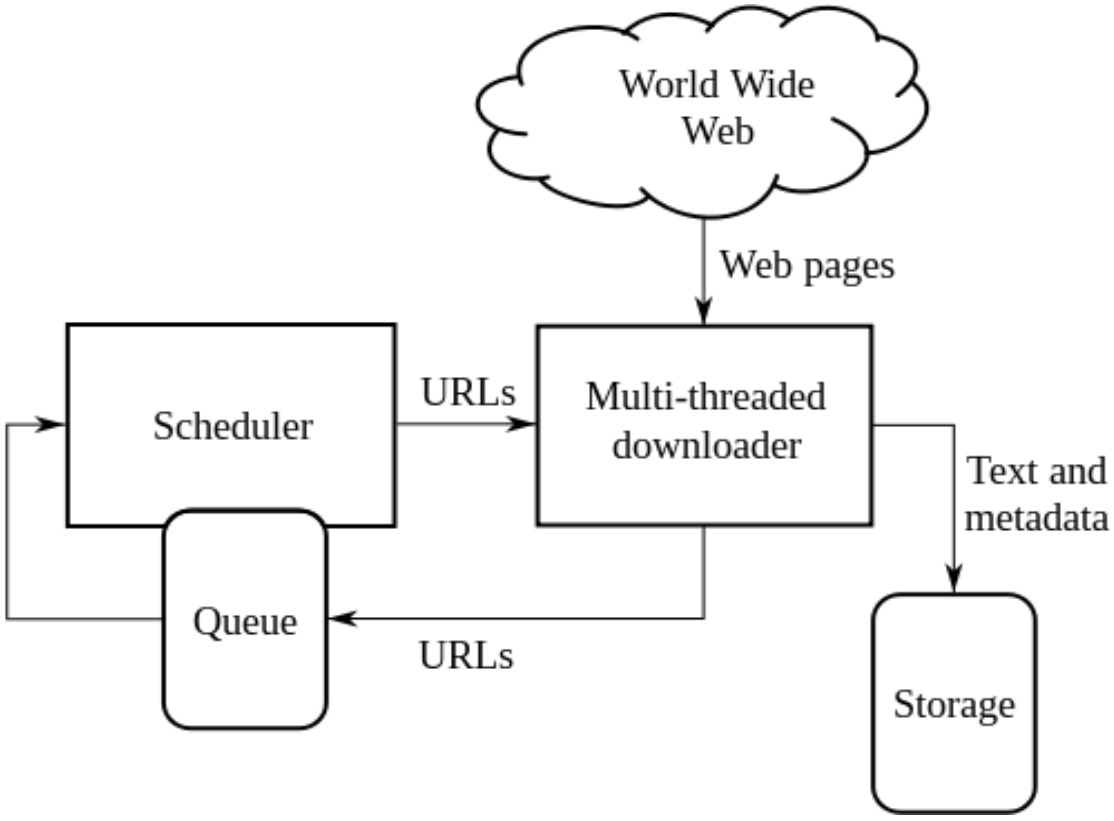
Web Search Engine and some other sites use Web crawling software to update their content by updating the novelty indexes of others sites' web content. Web crawlers can create a facsimile of all the pages they visit for later processing and updating the dynamic table optimizing the performance of the search engine resulting in whole riveting experience for the users.

Crawlers validate hyperlinks and HTML code.

1.2.1 CRAWLING POLICY

The web crawler comprises of four major steps:

1. A selection policy that states which pages to download,
2. A revisit policy that states when to check for changes to the pages,
3. A politeness policy that states how to avoid overloading web sites,
4. A parallelization policy that states how to coordinate distributed web crawlers.



1.3 WEB SCRAPING

Web scraping is a computer software technique of extracting information from websites.

Web Scraping software programs enable human interaction with the Internet by either executing low-level HTTP or by integrating it with a browser of any choice (say Chrome, Mozilla).

This method is a surmised aftereffect of Web Indexing utilizing a web bot or Web Crawler and is a strategy utilized via web indexes. But web scraping center's entirely on the development of raw data in HTML format into classified data of importance that might be stored and examined in a central repository or in table format.

Web scraping is identified with web mechanization and deals in reproducing user skimming by utilizing computer software. Technocrats who conduct web scraping mainly have the intention for online price comparison.

1.3.1 TECHNIQUES

HUMAN COPY-AND-PASTE: Some of the time even the best web-scraping technique can't supplant a user's manual survey and copy-paste and most of the times this can be the main desired and executable result when the sites that do web scraping intentionally set up obstacles to counteract machine mechanization.

TEXT GREPPING AND REGULAR EXPRESSION MATCHING: A herculean action to deracinate data from web pages can be based regex approaches.

HTTP PROGRAMMING: Date's from both dynamic and static web pages can be extracted all through the method of socket programming by beseeching HTTP query to the targeted web server.

HTML PARSERS: One a many sites have massive accumulations of webpages produced dynamically from an organized center (say database). Similar data that can be categorized under one umbrella will be regularly encoded into comparable webpages by a typical format. In data mining, a system which identifies this kind of formats in a specific data source, retrieves its content and makes a facsimile impression of it into a wrapper. Webpages of a wrapper induction framework adjust to a typical layout and which could be recognized using a URL common scheme which is the basis of any Wrapper Generation Algorithm.

DOM PARSING: By insertion of a web browser, e.g., Safari and Chrome, programs can segregate the dynamically generated content solely by using client side scripting. These browsers arrange webpages into a DOM tree, from which programs can extract and use the constituents for web scraping.

WEB SCRAPING SOFTWARE: Many software tools are accessible that could be utilized to alter results of web scraping. This tool is dedicated to automatic perceiving the data structure of a page or giving a recording interface that skips the step to write down web-scraping codes, or a few scripting functions that are required to delete and manipulate the contents and storing the scraped information in local or central databases.

VERTICAL AGGREGATION PLATFORMS: The stages for Vertical Aggregation Platforms make and check a large number of "bots" for particular verticals with no user inside the loop, and no work associated with a particular target website. This planning includes creating the knowledge base for the whole vertical platform and afterwards the stage instantiates the bots automatically. The stage's longevity is assessed by the nature of the information it derives and its adaptability or the speed of its surveying. So the above mentioned versatility is basically used to focus on the Long Tail of sites that regular accumulators discover confounded or excessively labor-intensive to collect content from.

SEMANTIC ANNOTATION RECOGNIZING: Usually the pages being scraped may grasp metadata or semantic markups and annotations, which could be utilized to place particular information bits. In the event that the footnotes are installed in the pages, this is probably a procedure of DOM parsing from which developers derive their content from. Alternatively the annotations, architected into a semantic layer, are put away and oversaw independently from the concerned web pages and eventually the scrapers can get the data classification from this specific layer before deleting the pages.

COMPUTER VISION WEB-PAGE ANALYZERS: This deals with techniques related to machine learning and computer vision whose objective is to recognize and collect data from web pages as per the requirements of the user.

1.4 TYPE OF SPAMMERS

Spammers are categorized into Flooders, Promoters, and Trojans.

A Flooder tags resources which as of recently exist in the system in a programmable manner.

A Promoter spams and concentrates on tagging his resources to push ubiquity, and considers other resources to be redundant. Their group participates with one another tagging every other resources ramifying to a strong connection heightening their positions and metamorphosing entire system according to their volition. He has a tendency to be the first to bookmark resources which draw in few followers if any.

A Trojan is refined and circumspect spammer. Its method is to copy and sham targeted users. He masks his pernicious plans by tagging effectively mainstream pages, however eventually he adds connections to his own particular reports which may be contaminated utilizing cross side scripting; the charlatanism emaciated from its side gradually taints the entire system.

CHAPTER 2

LITERATURE REVIEW

2.1 RELATED WORK

Expert recognizance generally includes building expert profiles by linking resources with the experts and utilizing IR methods on the profiles. Late methodologies include graph-based analysis of client systems in a group. There has been an algorithm focused around PageRank to process expert ranking of users on a social network. It has been proposed in different papers ranking users with fortification with the resources by dissecting different relationships in WWW.

There has been religious and pedantic discussions on detection and demoting of spammers so this could escalate the popularity of the sites and can also give the users a better experience.

It has been suggested that authenticity of users -their tags agree with those of the others – ought to be considered to handle a ranking of resources which is more impervious to spammers. There are likewise suggestions for distinguishing spammers in tagging systems focused around machine learning methodology.

We propose besides finding experts, nose diving the rank of the spammers in the ranked list of users can ameliorate the existing systems escalating their flexibility.

2.2 HITS ALGORITHM

Input: Adjacency matrix A of size m × m where m is total number of nodes in graph G and number of iterations k

Output: Authority and hub score vector $\vec{X}$ and $\vec{Y}$ respectively

1) $\vec{X}$=(1,1,1....,1) ∈ R^m
2) $\vec{Y}$=(1,1,1,...,1) ∈ R^m
3) While k
    a) For i=1 to m
        i) $X_j = \sum_{A(i,j)} Y_i$
        ii) End for
    b) For j=1 to m
        i) $Y_i = \sum_{A(i,j)} X_j$
        ii) End for
    c) Normalize $\vec{X}$
    d) Normalize $\vec{Y}$
4) End for
5) Return $\vec{X}$ and $\vec{Y}$

2.2.1 EXPLANATION

Here, the initial step is to extract the very important pages to query for searching. These belong to the crux lot and could be gotten by having first n pages resulted by a string-based search algorithm. Web pages in the base set and all hyperlinks around those pages structure a centered sub graph. The HITS calculation is performed just on this centered sub graph.

Scores of authority and hub are characterized as far as each other in a dependent recursion. An authority value is figured as the aggregate of the scaled hub values that indicate that page. A hub value is the aggregate of the scaled authority values of the pages it indicates. A few usages additionally think about the significance of the connected pages.

The algorithm executes iteratively till it converges according to the matrix Eigen vector, comprising of 2 general steps:

- Update Authority: every hub's Authority score to be equivalent to the summation of all of the Hub Scores of each one hub that indicates it.
- Update Hub: every Hub value is equivalent to summation of Authority Values of every one of the hub that is indicated.

HITS is evaluated on following lines:

- Initialization of both hub score vector and authority score vector to one.
- Loop the update authority and update hub rules.
- Normalization will be done by dividing each scores with the squared root of the summation of squared values of the scores.
- Repeat 2nd step for all hubs.

2.3 ROBUST PAGERANK

Robust page rank algorithm is an amelioration of the rudimentary page rank algorithm. The robust page rank is dependable if we have to fight against link spamming and link spammers ought to put all the more in purchasing new domain to build a system of ranking hub scores.

The Robust Page Rank of the node V can be composed equivalent to summation of commitment of different nodes to v.

$$pr_{\alpha}(v) = \sum_{u \in V(G)} ppr(u, v)$$

The robust PageRank diminish the impact of the most compelling hubs on the PageRank of a hub to metamorphise the ranking more vigorous against linkspamming. By diminishing impact of link spammers who get primary share of their PageRank from the connected graphicalnetwork, we expect that the link spammer lose an expansive segment of their PageRank. It is possible by diminishing the commitment of hubs with a commitment more than an threshold δ’ to δ’’.

$$Robustpr_{\alpha}^{\delta}(v) = \sum_{u \in V(G)} \min(ppr(u,v), \delta)$$

Since we diminish the commitment of a couple of most compelling hubs to the PageRank of a hub, we diminish the PageRank of spammed hubs more than the non-spammed hubs.

The calculation of a local-approximation of the Robust PageRank according to the following mathematical formulae. First, we can rewrite the Robust PageRank

$$Robustpr_{\alpha}^{\delta}(v) = \sum_{u \in V(G)} \min(ppr(u,v), \delta)$$

$$= \sum_{u \in V(G)} ppr(u,v) - \sum_{u \in S_{\delta}(v)} (ppr(u,v) - \delta)$$

$$= pr_{\alpha}(v) - \sum_{u \in S_{\delta}(v)} ppr(u,v) - \delta|S_{\delta}(v)|$$

2.4 WHO IS AN EXPERT

An expert is for the most part somebody with a large amount of information, technique or abilities in a particular domain. This is a direct indication and inference to the fact that good users are reliable and best sources for information.

2.4.1 RELATIONSHIP BETWEEN EXPERTISE AND QUALITY OF RESOURCE

Most rudimentary methodology of surveying the aptitude of a user in a context is quantified by how many times he has utilized the related tags on resources. It is the most prominent technique embraced by most collaborated systems today. Unfortunately, such a methodology inadvertently doesn't think about the actualities that amount is not proportional to quality, and spammers who randomly tag a supernumerary amount of resources may be taken as an expert (it is a misdemeanor).

According to collaborative tagging, users tags to resources to channel the extraction of resources that are handy and required later. Thus, we accept that a expert ought to be somebody who not just has a substantial accumulation of resources attached with a specific tag, yet has a tendency to add rich resources to their accumulations.

Akin to HITS algorithm, authority and hubs commonly fortify one another. The contrasts for our situation are that tagging includes two various types of interrelated elements, to be specific users and resources, and that no one but users can indicate resources however not the other way around. Along these lines for our situation users will

just accept hub scores though resources will just get authority scores. This bodes well in light of the fact that experts demonstration as hubs when we discover valuable resources using users, and resources go about as authority.

Common fortification between users and resources have been examined we are quite skeptical if this much is only required in evaluation. Users generally recognize newly found resources after different users tagged them. As such, there is a colossal amount of risk that users gain from one another as opposed to finding data without anyone else present as in performing a Web search. Consequently the second assumption follows that resources are authorities.

2.4.2 TYPES OF EXPERTS

A veteran is a user who tags a larger number of resources than the normal user. He has a tendency to be around the first user to tag resources which will in the long run get to be truly mainstream inside the network. Henceforth, he is an identifier with numerous followers.

A newcomer is an approaching expert who is just here and there around the first to "uncover" a resource. More often than not, the resources are popular when he tags them.

A geek will be comparative to a veteran however has fundamentally a larger number of tags than veteran.

2.5 FOLLOWER VS. DISCOVERER

According to HITS algorithm, 2 users have the exact expertise score despite the fact that one is first to tag some resources and alternate is essentially tagging the resources in light of the fact that they are now mainstream in the network. What's more, a spammer who needs elevate some Web pages to different users can without much of a stretch endeavor this shortcoming and help his expertise score by tagging heaps of famous resources.

Experts ought to be the identifiers of riveting resources, as opposed to the followers who discover the same resources later in light of fact that the resources have gotten well known as of recently. Genuinely talking the prior a user tagged a resource, the greater score he deserves to receive. So the rationality of taking time as a consideration for deciding the expert scores of a user. The time stamp of the tagged resource is a rational consideration of that he is so critical to novelty in regard the context.

The follower discovers supposition is a sensible and an alluring in light of the fact that experts ought to be the users who brought great resources to the consideration of learners. So this transforms our strategy of expertise ranking, cocooned against different type of spammers.

CHAPTER 3

PROPOSED ALGORITHM

3. ALGORITHM

Input: [M, N, B, C, K], Where M= Total number of users, N= Total number of resources, K= Total number of iterations, B= Set of bookmarks, C= Credit Function

Output: [A, B], Where A= Ranked list of users, B= Ranked list of resources

1. $\vec{E}$=(1,1,1....,1) ∈ Q^M
2. $\vec{R}$=(1,1,1,...,1) ∈ Q^N
3. A ←Adjacency Matrix (B, C)
4. For j=1 to K
    a. $\vec{E}$= $\vec{R}$ × A^T
    b. $\vec{R}$ = $\vec{E}$ × A
    c. Normalize $\vec{E}$
    d. Normalize $\vec{R}$
5. End for
6. X ← List of users sorted according to their expertise score in $\vec{E}$ return X
7. Y ← List of resources sorted according to their quality score in $\vec{R}$ return Y

3.1 EXPLANATION

The project's scheme is to use the new algorithm for ranking users in a collaborative tagging system by taking experts specified previously. First assumption of experts includes the schema of dexterity of users and the idiosyncrasy of the resources commonly fortifying one another. We characterize $\vec{E}$ as a vector of expertise scores of users: $\vec{E}$ = (u1,u2,....) where M = |E| is the number of unique users in Rt. $\vec{R}$ is a vector of quality scores of resources: $\vec{R}$ = (d1,d2,...) where N = |R| is the number of unique resources in Rt. $\vec{E}$ and $\vec{R}$ are instated by setting each component to 1. Essentially, the definite worth of the components could be subjective as long as they are all equivalent, as the vectors will be standardized in later operations.

Common fortification alludes to the thought that the users expertise score relies upon the resources quality score to which he tagged t, and the resources quality score relies upon user's expertise score who tagged t to it. we set up a adjacency matrix A of size M × N where A(i,j) := 1 if user i has tagged t to resource j, and A(i,j) := 0 overall.

$$\overrightarrow{E_k} = \boldsymbol{\alpha_k} \boldsymbol{A^T} \boldsymbol{\vec{R}_{k-1}}$$

$$\overrightarrow{R_k} = \boldsymbol{\beta_k} \boldsymbol{A} \boldsymbol{\vec{E}_{k-1}}$$

To execute the very notion of discoverers and followers, we set up the adjacency matrix A in a manner not quite the same as the above system for allocating either zero or one. We applied the power method in accord with the following mathematical statement to generate the adjacency matrix A:

$$A_{i,j} = \left|\left\{u \middle| (u,t,d_j,c),(u_i,t,d_j,c_i) \in R_t \wedge c_i < c\right\}\right| + 1$$

As indicated by above mathematical statement, A (i, j) is equivalent to one or more the amount of users who have tagged t to resource d_j after u_i. Thus, if u_i is start user to dole out t to d_j, A (i, j) will be equivalent to the aggregation of users who have assigned t to d_j. On the off chance that u_i is the latest user to dole out t to d_j, A (i, j) will be equivalent to one. The impact of this instantiation of matrix A is that we have a classified timetable of all users who tagged a given resource d_j. Then the final step is to assign a fitting credit function value to users by application of credit function C to A i.e, $A_{i,j} = C(A_{i,j})$.

If we contemplate for a linear function C(x) =x but it may not be the most pertinent one.

Along these lines, when the expertise scores are ascertained by power method, users who have tagged a resource at first will guarantee a greater amount of quality score than that of the individuals who have tagged the resource later. One worry of credit function is, that explorers, who have tagged a resource at first will accept nearly a grater expertise score despite the fact that they may have not helped any viable resource from there on.

We conjecture that the model of legitimate credit function C is ought to be an increasing function concave downward that is $\frac{dC(x)}{dx} > 0$ and $\frac{d^2C(x)}{dx^2} <= 0$. We expect that the discoverer of a document is ought to get a score greater than the followers however it is necessary to lessen the disparity between the scores because if a user is finding a new document initially and it is a popular tag later so he is bound to get high scores. But think

he is sitting ideally after that tagging so his score must not increase at that pace .So a decreasing first derivative is needed. In our algorithm, we lead our tries with $C(x):= x^y$, where $x>0$ and $0<y<1$.

CHAPTER 4

EXPERIMENTS AND EVALUATION

4.1 DATASET AND METHODOLOGIES

The main juggernaut for evaluating the algorithm was the lack of real time data set. We then scrape the site Delicious.com and collected the data of users, tags, documents and date and time of tagging.

A bookmark incorporates the Delicious userid of a user, details written in bookmark, attached tag, and its time of creation. Because of limitations forced by Delicious, we recovered up to a most extreme of 8 client bookmarks for every URL.

We also collected scraped and simulated data of this delicious.com site from different web mining enthusiasts over online beseeching.

The data which we collected had two types of user profiles both experts and spammers. Experts are of three type's geeks, veterans and novice, and spammers are flooders, promoters and Trojans.

4.2 BEHAVIOUR

We astutely observed the execution of our calculation by scrupulous comparison that comes about from our algorithm and those returned by HITS and robust page rank calculation that is focused around the amount of tagging.

Applying the two algorithms on the users and resources in the four chose data sets, we acquire all results and graphs, which demonstrates the ramifying normalized expertise score graphs. It hints our algorithm differentiates separated qualities than HITS and Page Rank, for example, the contrast in expertise scores in both the algorithms is more than in HITS and page rank.

An alternate discovering will be the staircase-like shape of robust PageRank brought about by the number recurrence depends on which it is based. This methods page rank has a tendency to gathering users into basins of equivalent expertise score as opposed to appointing a singular rank to every user.

4.3 PROMOTING EXPERTS

To explore how distinctive experts are positioned by our robust computation, we create, 4 data sets having 2981 users in every data set utilizing xml parsing. Then we applied our calculation, firstly HITS calculation and then PageRank to those data sets. We watch that the significant distinction between our calculation and the other calculations is predictable around all the 4 data sets. In our algorithm, geeks are for the most part positioned above veterans, thus positioned above novices.

4.4 DEMOTING SPAMMERS

Robust page rank performs most noticeably awful in catching spammers in a collaborative tagging or online bookmarking website as it predominantly centers on whom the users will be tagging and how numerous are emulating the users. The misleading of all people is not pondered about seriously while figuring of these positions and the spammers are inadvertently given higher positions.

HITS perform marginally as it has a tendency to nose dive promoters to low positions, despite the fact that is not ready to plummet flooders and Trojans. Unluckily, flooders specifically are regularly found in collaborative tagging systems.

Our calculation executes around the three calculations. First, it effectively plunges both the spammers that are flooders and promoters, and in each case it crashes their ranks of the spammers than HITS and robust page rank. Besides, our calculation is likewise

equipped to downgrade the spammer like Trojans who utilize a significantly more complex plan. Though they are still positioned at a higher rank than the other two, but no Trojans were seen in the range of ranks where there were actual experts. In the experimentation which we did the top 40 experts ought to be the users who got us bedazzled with the richness in their tagging's.

It also shows the present structure as of recently performs sensibly well in getting a good riddance of Trojans in our pertinent reach. In certainty, the juggernaut with Trojans is that it is a conundrum to downgrade their ranks amid not plummeting good users concurrently. As from rationality, a Trojan is a rich arsenal of resources. Users encroaching resources in a Trojan's repository is needed to validate the quality score of the resources. It is mathematically calculated by our algorithm, to decide if they meet the dejure and defacto, and rich resources before having a tryst with them.

See Trojan is an imposter of an expert. So detection of Trojan just by this approach in a cent percent successful way is quite difficult.

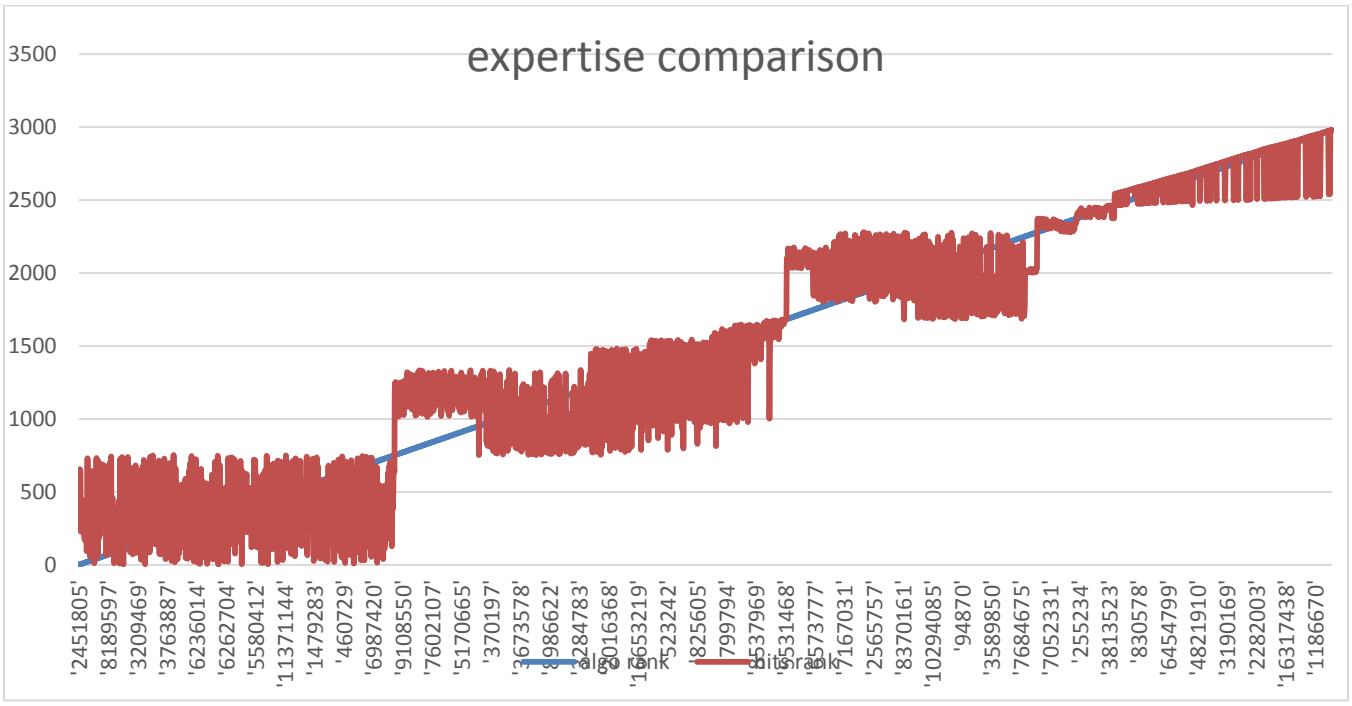
But sometimes we may hit the jackpot due to some preposterous decision of the spammers.

Recent studies show that there is only 5% quality data in any arena we consider. Spammers mainly form a very core unit.

If after using our algorithm and getting the expert ranks we form a graph with edge capacities difference between the expert scores and there is an edge between two experts if they have tagged same documents, then we find the SCC, maximal clique and find min cut.

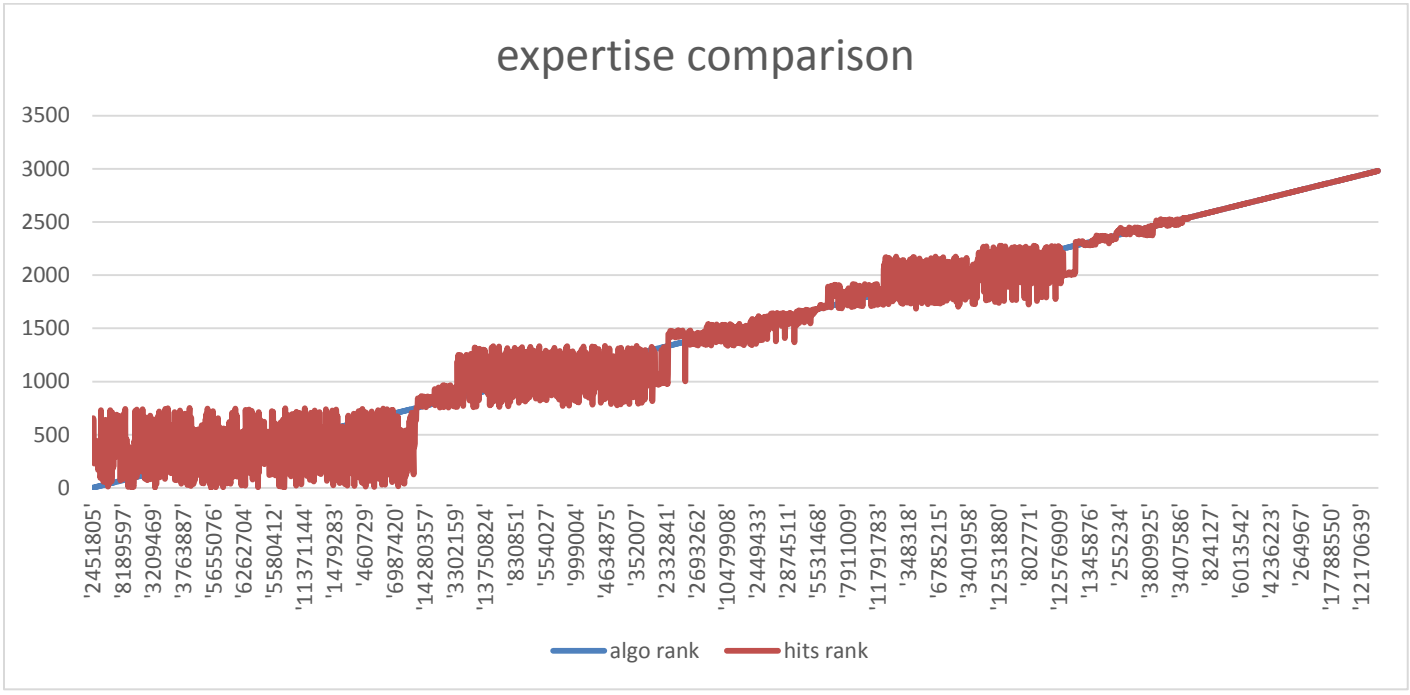
We can vandalize the spamming links and enervate their links. But due to time constraints we are not able to perform on this line. But surely we can add this to the future work if we consider it later.

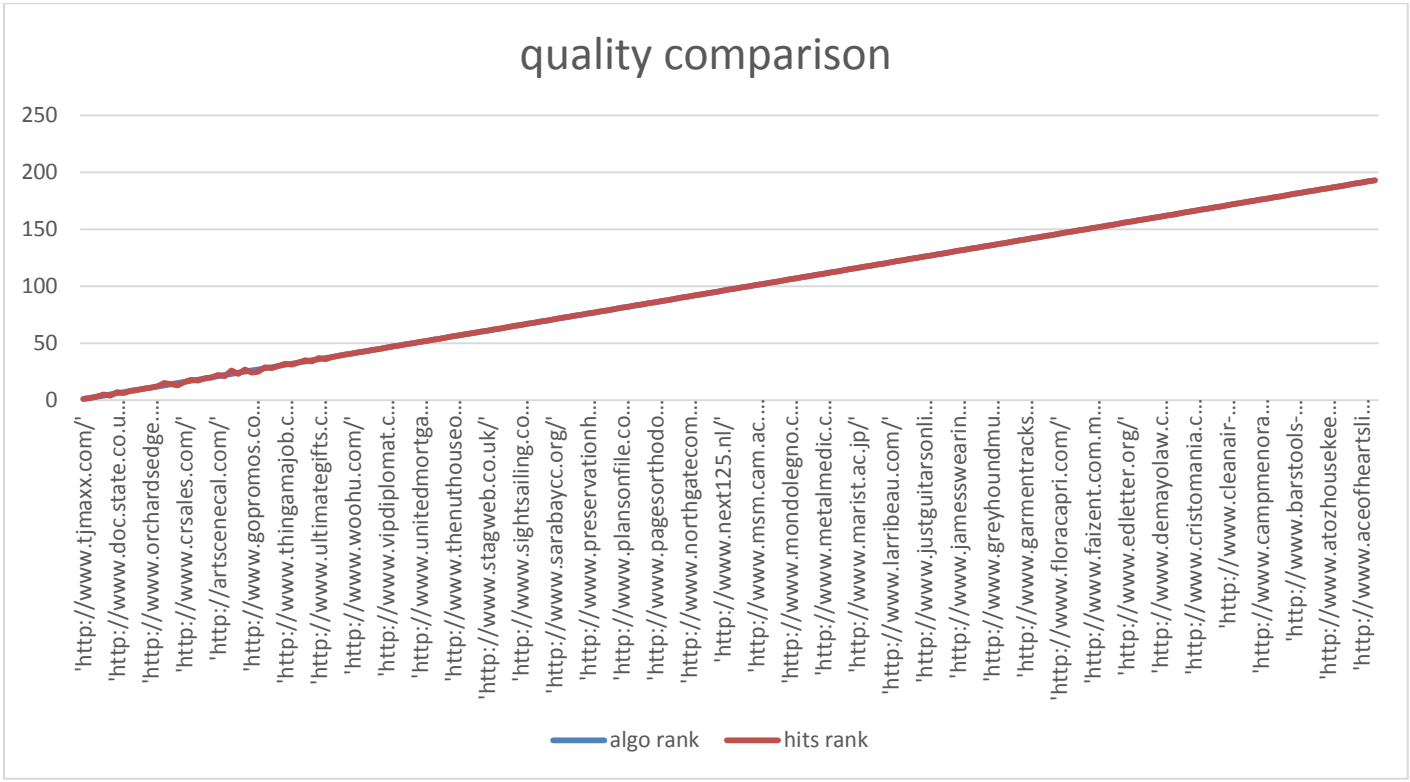
$C(x) := x^{0.5}$


expertise comparison
3500
3000
2500
2000
1500
1000
500
0
algo rank
hits rank

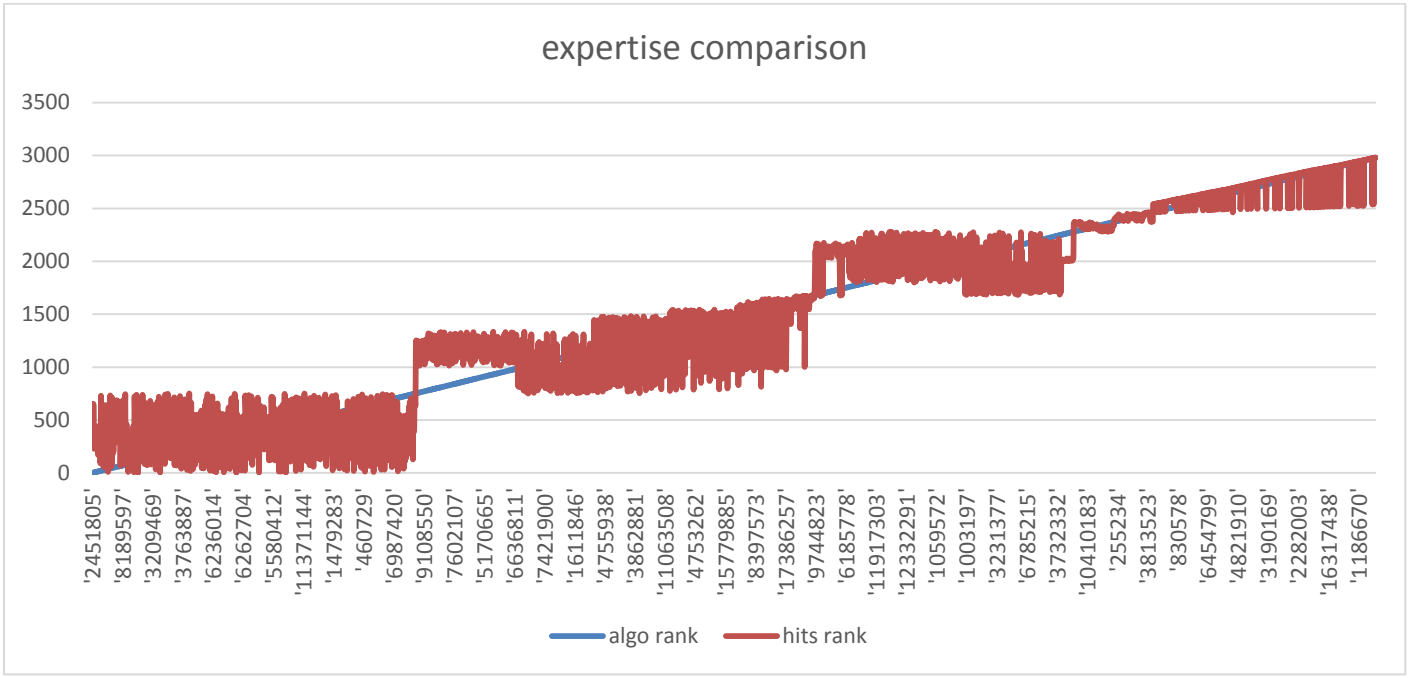

quality comparison
250
200
150
100
50
0
algo rank
hits rank

$C(x) := x^{0.1}$


expertise comparison
3500
3000
2500
2000
1500
1000
500
0
'2451805'
'8189597'
'3209469'
'3763887'
'5655076'
'6262704'
'5580412'
'11371144'
'1479283'
'460729'
'6987420'
'14280357'
'3302159'
'13750824'
'830851'
'554027'
'999004'
'4634875'
'352007'
'2332841'
'2693262'
'10479908'
'2449433'
'2874511'
'5531468'
'7911009'
'11791783'
'348318'
'6785215'
'3401958'
'12531880'
'802771'
'12576909'
'1345876'
'255234'
'3809925'
'3407586'
'824127'
'6013542'
'4236223'
'264967'
'17788550'
'12170639'
algo rank
hits rank


quality comparison
250
200
150
100
50
0
'http://www.tjmaxx.com/'
'http://www.doc.state.co.u...'
'http://www.orchardsedge....'
'http://www.crsales.com/'
'http://artscenecal.com/'
'http://www.gopromos.co...'
'http://www.thingamajob.c...'
'http://www.ultimategifts.c...'
'http://www.woohu.com/'
'http://www.vipdiplomat.c...'
'http://www.unitedmortga...'
'http://www.thenuthouseo...'
'http://www.stagweb.co.uk/'
'http://www.sightsailing.co...'
'http://www.sarabaycc.org/'
'http://www.preservationh...'
'http://www.plansonfile.co...'
'http://www.pagesorthodo...'
'http://www.northgatecom...'
'http://www.next125.nl/'
'http://www.msm.cam.ac....'
'http://www.mondolegno.c...'
'http://www.metalmedic.c...'
'http://www.marist.ac.jp/'
'http://www.larribeau.com/'
'http://www.justguitarsonli...'
'http://www.jamesswearin...'
'http://www.greyhoundmu...'
'http://www.garmentracks...'
'http://www.floracapri.com/'
'http://www.faizent.com.m...'
'http://www.edletter.org/'
'http://www.demayolaw.c...'
'http://www.cristomania.c...'
'http://www.cleanair-...'
'http://www.campmenora...'
'http://www.barstools-...'
'http://www.atozhousekee...'
'http://www.aceofheartsli...'
algo rank
hits rank

$C(x) := x^{0.9}$


expertise comparison
3500
3000
2500
2000
1500
1000
500
0
'2451805'
'8189597'
'3209469'
'3763887'
'6236014'
'626704'
'5580412'
'11371144'
'1479283'
'460729'
'6987420'
'9108550'
'7602107'
'5170665'
'6636811'
'7421900'
'1611846'
'4755938'
'3862881'
'11063508'
'4753262'
'15779885'
'8397573'
'17386257'
'9744823'
'6185778'
'11917303'
'12332291'
'1059572'
'1003197'
'3231377'
'6785215'
'3732332'
'10410183'
'255234'
'3813523'
'830578'
'6454799'
'4821910'
'3190169'
'2282003'
'16317438'
'1186670'
algo rank
hits rank


quality comparison
250
200
150
100
50
0
'http://www.tjmaxx.com/'
'http://www.metroweekly.com/'
'http://www.orchardsedge.com/'
'http://www.crsales.com/'
'http://artscenecal.com/'
'http://www.gopromos.com/'
'http://www.hoopscooponline.c...
'http://www.nuwa.com/'
'http://www.willowcreekct.com/'
'http://www.veritysystems.com/'
'http://www.travelgrenada.com/'
'http://www.tannery.com/'
'http://www.southsidejohnny.c...
'http://www.sfosnow.com/'
'http://www.rumpsville.com/'
'http://www.ppcwm.org/'
'http://www.pharmaceuticalsal...
'http://www.oxyvinyls.com/'
'http://www.norcalweather.co...
'http://www.neurosis.com/'
'http://www.mousselieres.com/'
'http://www.mondolegno.com/'
'http://www.metalmedic.com/'
'http://www.marist.ac.jp/'
'http://www.larribeau.com/'
'http://www.justguitarsonline.c...
'http://www.jamesswearingen....
'http://www.greyhoundmuzzles...
'http://www.garmentracksetc.c...
'http://www.floracapri.com/'
'http://www.faizent.com.my/'
'http://www.edletter.org/'
'http://www.demayolaw.com/'
'http://www.cristomania.com/'
'http://www.cleanair-...
'http://www.campmenorah.co...
'http://www.barstools-and-...
'http://www.atozhousekeeping....
'http://www.aceofheartslimo.c...
algo rank
hits rank

CHAPTER 5

CONCLUSION

5. CONCLUSION

We proposed an algorithm for ranking experts in a collaborative tagging system and compared the data and ranking of several users by scrapping the data from the coveted Delicious.com site and embezzle these scrapping results in our study and examination of our algorithms. Our evaluation alludes what we proposed on the lines of existing lines just escalating the dimension of review; is more capacitate enough at differentiating a whole gamut of experts and is more impervious to myriad of spammers than the HITS and robust Page Rank algorithms.

The point that was conspicuously noted down is that the proposed algorithm evaluated and calculates virtuosity based on user's dexterity to discover novelty which is rich and resourceful but it is of course only one dimension of his dexterity and expertise. In addition a crux objective of the collaborative tagging systems is to discover rich and riveting resources and documents such that the virtuosity dimension of the users are fully and religiously examined by different algorithms is important for such systems.

We long our pedantic examination sprout out a slew number of research directions. We have mentioned just after the results

This algorithm can be ameliorated, for instance the graph concepts perfunctorily touched. We will thus use more data analysis to explore more spamming possibilities and find more ways to tackle and collar them. One such method is Machine learning.

REFERENCES

1. Telling Experts from Spammers:Expertise Ranking in Folksonomies SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA.Copyright 2009 ACM 978-1-60558-483-6/09/07

2. P. Heymann, G. Koutrika, and H. Garcia-Molina.Fighting spam on social web sites: A survey ofapproaches and future challenges. IEEE Internet Computing, 11(6):36{45, 2007}

3. M. G. Noll and C. Meinel. Exploring social annotations for web document classication. In Proc. of ACM Symposium on Applied Computing, pages 2315{2320, Fortaleza, Brazil, 2008

4. Madkour, T. Hefni, A. Hefny, and K. S. Refaat.Using semantic features to detect spamming in social bookmarking systems. In Proc. of ECML PKDD Discovery Challenge Workshop, Belgium, 2008

5. S. Golder and B. A. Huberman. Usage patterns of collaborative tagging systems. Journal of Information Science, 32(2):198–208, April 2006.

6. P. Heymann, G. Koutrika, and H. Garcia-Molina. Can social bookmarking improve web search In Proc. of 1st ACM Int'l Conf. on Web Search and Data Mining (WSDM'08), pages 195–206.ACM, February 2008

7. B. Krause, C. Schmitz, A. Hotho, and G. Stumme. The anti-social tagger:detecting spam in social bookmarking systems.